Learning Diverse Bimanual Dexterous Manipulation Skills from Human Demonstrations

Bohan Zhou, Haoqi Yuan, Yuhui Fu, Zongqing Lu

01/22/2026

Motivation

PEKING UNIVERSITY

Bimanual manipulation is fundamental for everyday tasks.

- **“symmetry" collaborative tasks** (*e.g.* lifting a heavy box)
- **“asymmetry” tasks** (*e.g.* twisting a bottle cap)

Bimanual robotic dexterous manipulation is largely unexplored.

- **High DOF** (*e.g.* Shadow Dexterous Hand DOF=5+4*4+6=27)
- **Coordination** (*e.g.* wringing towels, sewing clothes, playing the piano, tying shoelaces, assembling parts)
- **Few Benchmarks** (*e.g.* Bi-DexHands [1], more focus on single hand / grippers)

We would ask:

“Can we learn diverse bimanual dexterous manipulation skills in a unified and scalable way?”

[1] Chen, Yuanpei, et al. Towards human-level bimanual dexterous manipulation with reinforcement learning. NeurIPS 2022.

Introduction

Our answer:

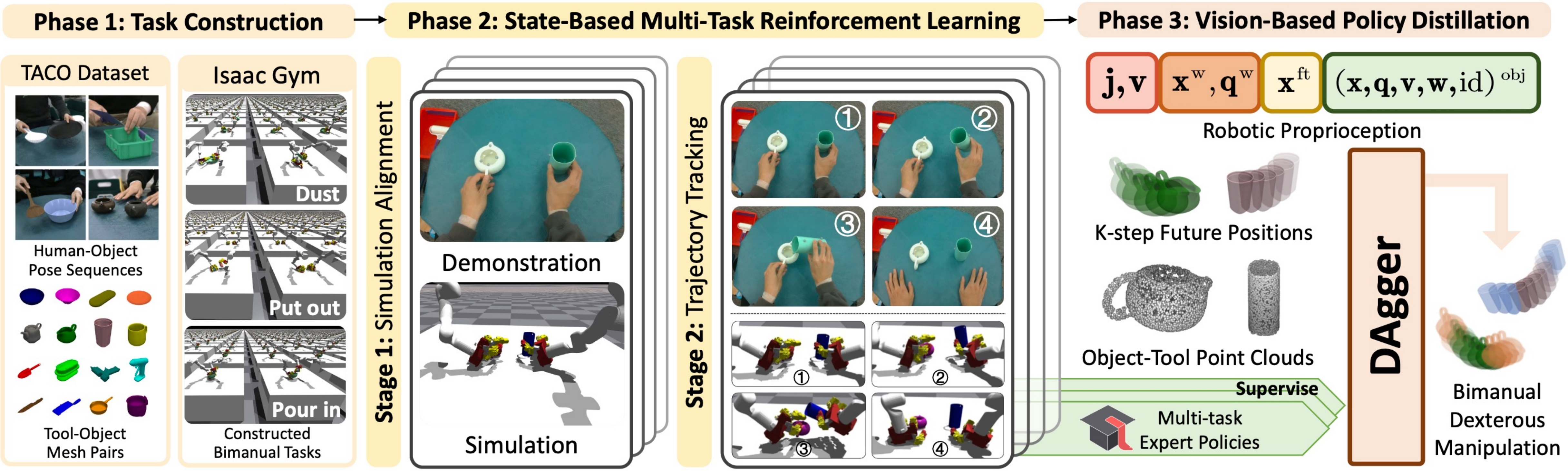
A unified and scalable framework, generally learning diverse bimanual dexterous manipulation skills from human demonstrations

Feature	BiDexHD	Related Work
Task Design	Automatically construct diverse bimanual tasks from human demonstrations	Focus on existing benchmarks or a limited range of tasks
Solution	Generally-designed two-stage reward function	Task-specific reward engineering

BiDexHD is evaluated across **141** constructed tool-using tasks over **6** categories from **TACO** dataset and **11** collaborative tasks from **ARCTIC** dataset, demonstrating zero-shot capabilities and scalability.

Methodology

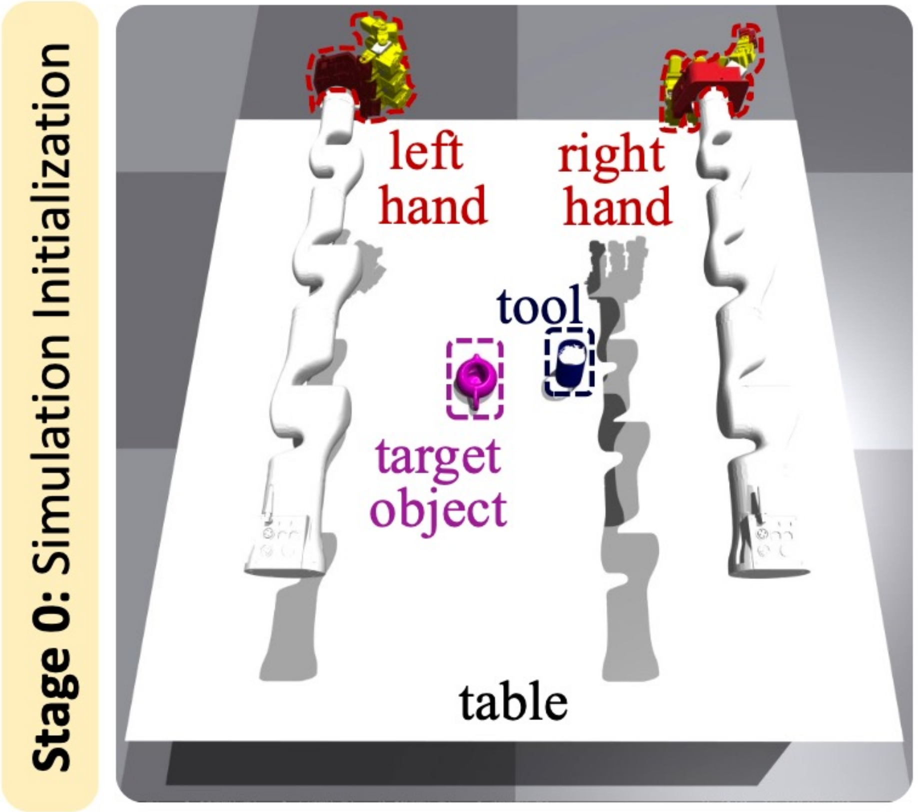
Phase 1: Constructing individual bimanual tasks from human demonstrations parallelly
Phase 2: Learning diverse state-based policies via multi-task reinforcement learning
Phase 3: Distill a group of learned policies into a vision-based policy for deployment



Phase 1: Task Construction

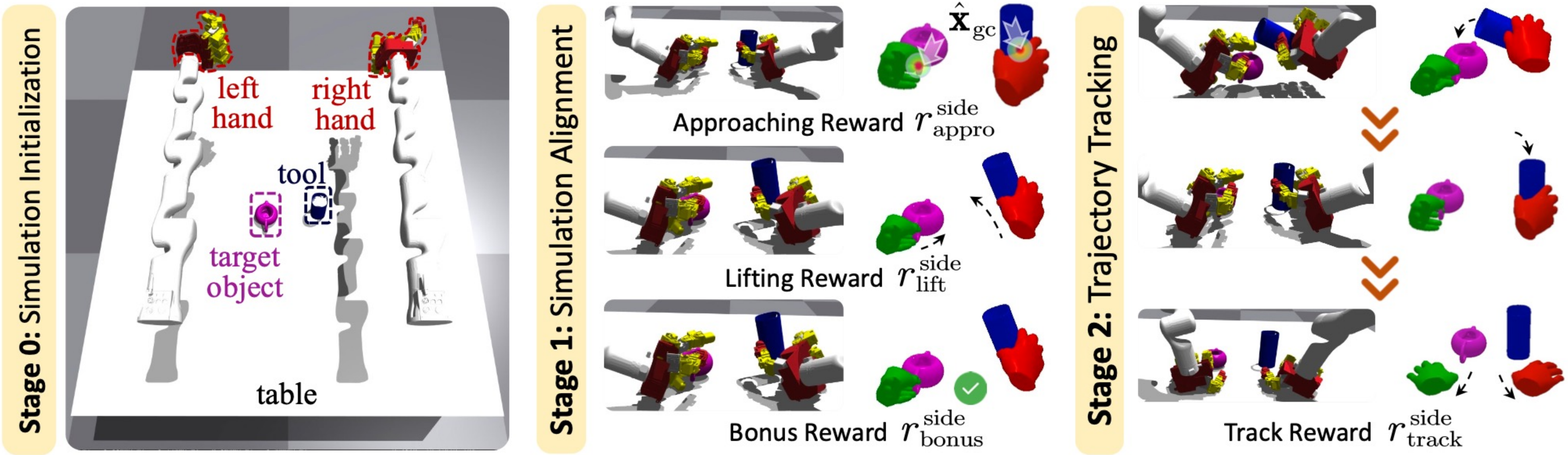
北京大學 PEKING UNIVERSITY

- Initial Joint angles & velocities: **zero**. (States of wrist and fingertips are calculated from forward kinematics.)
- Initial object poses: sampled from a fixed **Gaussian distribution**.
- Hand Joint angles: optimized from human fingertip positions via **AnyTeleop**.
- Arm joint angles: calculated from human wrist pose via **IK** based on the robot's palm base pose.



Phase 2: Multi-Task RL

- Target: learning a state-based policy for tasks that require similar behaviors via **multi-task reinforcement learning**
- Insight: **object-centric**, generalizable skill learning from the **object poses**, without additional **pre-grasp poses** estimated upon manipulated objects



Phase 2: Multi-Task RL

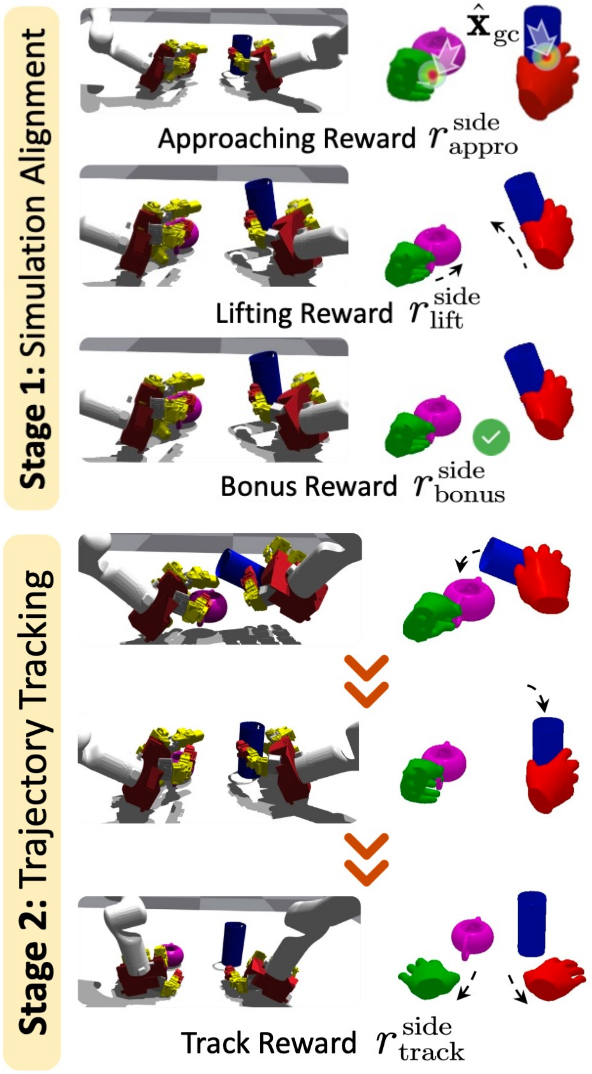
Stage 1. Simulation Alignment

Align the state of simulation to the first step in a trajectory by moving the tool or the target object from the initial pose to $\boldsymbol{\tau_0}$

- The left hand: approach, grasp or stabilize the target object
- The right hand: approach, grasp and hold the tool
- Success Detection: both objects reach the specified pose for a **sustained u-step duration**

Stage 2. Trajectory Tracking

Both hands are expected to maintain their hold and **ensure the objects follow their pre-defined trajectories** from the human demonstration dataset to perform the manipulations in sync.



Phase 2: Simulation Alignment

北京大學 PEKING UNIVERSITY

1. **Approaching Reward**: encourages both dexterous hands to approach and remain near their grasp centers

$$r_{\text{appro}}^{\text{side}} = -\|\mathbf{x}_t^{\text{side,w}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2 - w_r \sum^{m} \|\mathbf{x}_t^{\text{side,ft}} - \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}}\|_2$$

$$\text{where} \quad \hat{\mathbf{x}}_{\text{gc}}^{\text{obj}} = \frac{1}{L} \sum \text{NN}\left(\mathcal{P}, L, \frac{\hat{\mathbf{x}}_0^{\text{side,w}} + \sum^{m} \hat{\mathbf{x}}_0^{\text{side,ft}}}{m+1}\right)$$

Grasp center: first compute the mean of wrist & fingertip position at τ_0 as an **anchor**, then uniformly sample 1,024 surface points from the object mesh and take the **centroid of the top-L nearest** samples

Phase 2: Simulation Alignment

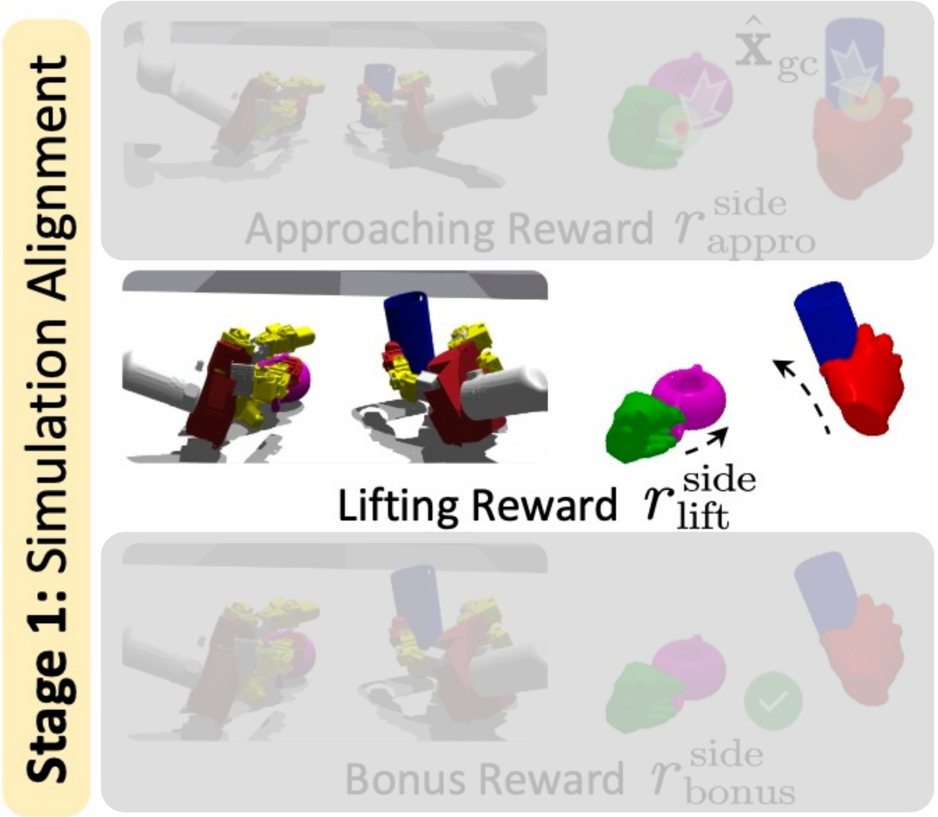
北京大學 PEKING UNIVERSITY

2. **Lifting Reward:** encourages holding objects tightly in hands and lifting to desired reference poses

$$r^{\text{side}}_{\text{pos}} = \max\left(1 - \frac{\|\mathbf{x}^{\text{obj}}_t - \hat{\mathbf{x}}^{\text{obj}}_0\|_2}{\|\mathbf{x}^{\text{obj}}_0 - \hat{\mathbf{x}}^{\text{obj}}_0\|_2}, 0\right)$$

$$r^{\text{side}}_{\text{quat}} = -\mathbb{D}_{\text{quat}}\left(\mathbf{q}^{\text{obj}}_t, \hat{\mathbf{q}}^{\text{obj}}_0\right)$$

$$r^{\text{side}}_{\text{lift}} = \left(r^{\text{side}}_{\text{pos}} + w_q r^{\text{side}}_{\text{quat}}\right) \cdot \mathbb{I}\left(\|\mathbf{x}^{\text{side,w}}_t - \hat{\mathbf{x}}^{\text{obj}}_{\text{gc}}\|_2 \le \lambda_{\text{w}}\right) \cdot \mathbb{I}\Big(\sum^{m}\|\mathbf{x}^{\text{side,ft}}_t - \hat{\mathbf{x}}^{\text{obj}}_{\text{gc}}\|_2 \le \lambda_{\text{ft}}\Big)$$



Phase 2: Simulation Alignment

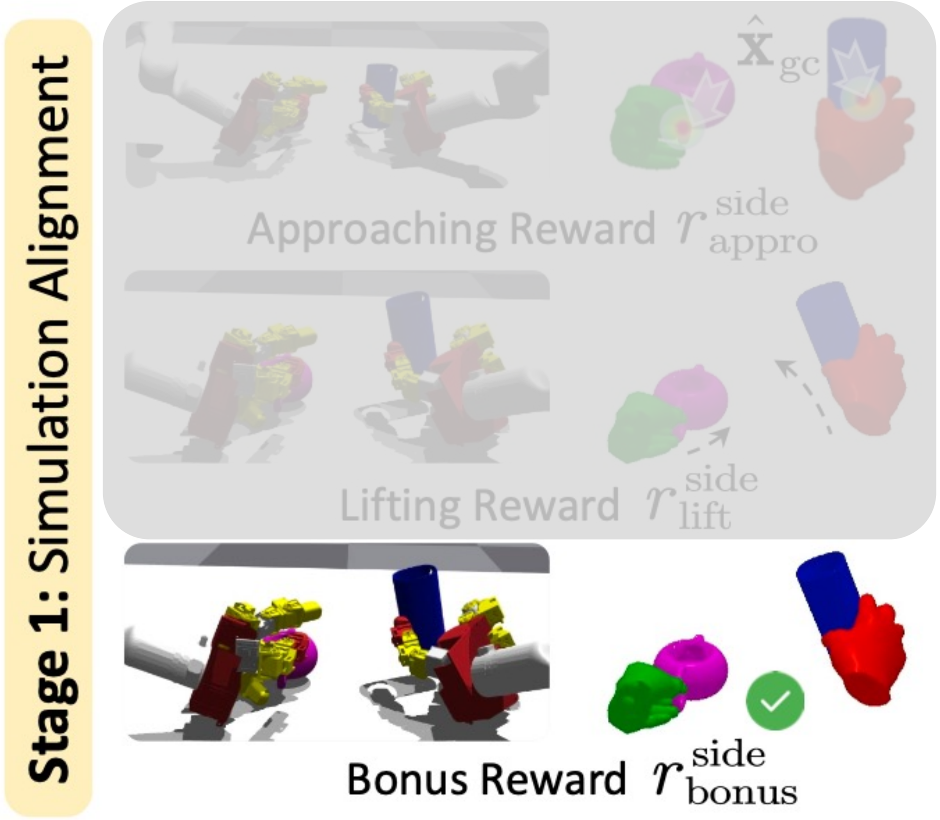
3. **Bonus Reward**: incentivizes the target object or the tool to keep staying at their reference poses

$$r^{\text{side}}_{\text{bonus}} = \begin{cases} \frac{1}{1+\|\mathbf{x}^{\text{obj}}_t - \hat{\mathbf{x}}^{\text{obj}}_0\|_2} & \text{if } \mathbb{I}\left(\|\mathbf{x}^{\text{obj}}_t - \hat{\mathbf{x}}^{\text{obj}}_0\|_2 \leq \varepsilon_{\text{succ}}\right) \\ 0 & \text{otherwise.} \end{cases}$$

Stage one is considered successful only if both $r^{\text{left}}_{\text{bonus}}$ & $r^{\text{right}}_{\text{bonus}}$ are positive for at least u consecutive steps

The total alignment reward is the linear weighted sum

$$r^{\text{side}}_{\text{align}} = w_1 r^{\text{side}}_{\text{appro}} + w_2 r^{\text{side}}_{\text{lift}} + w_3 r^{\text{side}}_{\text{bonus}}$$



Phase 2: Trajectory Tracking

北京大學 PEKING UNIVERSITY

Tracking Reward: encourages the dexterous hands to precisely track the desired positions at each timestep in a trajectory starting from the reference timestep

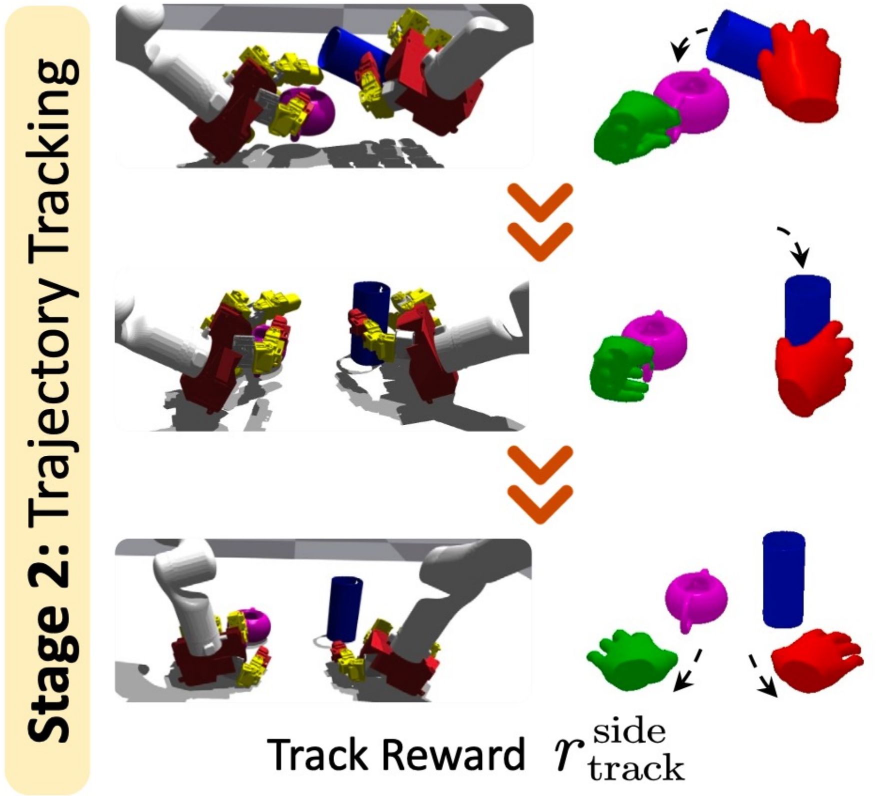
$$r_{\text{track}}^{\text{side}} = \begin{cases} \exp\left(-w_t \|\mathbf{x}_{t_i}^{\text{obj}} - \hat{\mathbf{x}}_i^{\text{obj}}\|_2\right) & \text{if stage 1 succeeds} \\ 0 & \text{otherwise.} \end{cases}$$

Tracking frequency *f* addresses human-robot gap:

$$i = \lceil t_i / f \rceil \in [0, l)$$

We adopt **IPPO** to learn a unified policy from:

$$r_{\text{total}}^{\text{side}} = r_{\text{align}}^{\text{side}} + w_4 r_{\text{track}}^{\text{side}}$$



Phase 3: Policy Distillation

Under the supervision of a group of state-based teacher policies, for each task category $\nu \in V$, We employ **Dagger** to distill vision-based $\pi_\phi^{\text{side}}(\mathbf{a}_t^{\text{side}}|\mathbf{o}_t^{\text{side}}, \mathbf{p}_t^{\text{side}}, \mathbf{a}_{t-1}^{\text{side}})$

- Object Pose → Object PointClouds $\mathbf{p}_t^{\text{side}} \in \mathbb{R}^{K\times 3}$

> **Pre-sample** 4,096 surface points per object mesh; at each timestep, draw a **subset**, **transform** them by the current object pose, and add **Gaussian noise** for robustness

- Future Object Positions $\mathrm{pc}_t^{\text{obj}} \in \mathbb{R}^{P\times 3}$

> It incorporates more information about the motion of objects (*e.g.* **movement direction and speed**) in the near future, facilitating zero-shot transfer

K	Train m_1	Train m_2	Test Comb m_1	Test Comb m_2	Test New m_1	Test New m_2
0	98.01	72.09	94.36	46.64	93.96	49.27
5	99.38	74.59	92.85	48.43	94.79	53.71

Experiments: Configuration

TACO [1] is a large-scale bimanual tool-using dataset.

- **6** categories = {Dust, Empty, Pour in some, Put out, Skim off, Smear}
- **141** human demonstrations, **80%** for training, **20%** for testing
  - **Test Comb**: objects in the training set, different behavior
  - **Test New**: target object or tool not in the training set

Metrics

$$\mathbb{I}_1 : \ \exists 0 < t < T - u \sum_{t}^{t+u} \prod^{\{\text{tool,object}\}} \mathbb{I}\left(\|\mathbf{x}_t^{\text{obj}} - \hat{\mathbf{x}}_0^{\text{obj}}\|_2 \le \varepsilon_{\text{succ}}\right) \cdot \mathbb{I}\left(\mathbb{D}_{\text{quat}}(\mathbf{q}_t^{\text{obj}}, \hat{\mathbf{q}}_0^{\text{obj}}) \le \varepsilon_{\text{succ}}\right) = u$$

$$\mathrm{m}_2 = \frac{1}{nl} \sum^{n} \sum_{i=0}^{l-1} \prod^{\{\text{tool,object}\}} \mathbb{I}\left(\|\mathbf{x}_{t_i}^{\text{obj}} - \mathbf{x}_i^{\text{obj}}\|_2 \le \varepsilon_{\text{track}}\right) \cdot \mathbb{I}\left(\mathbb{D}_{\text{quat}}(\mathbf{q}_{t_i}^{\text{obj}}, \mathbf{q}_i^{\text{obj}}) \le \varepsilon_{\text{track}}\right)$$

[1] Liu, Yun, et al. Taco: Benchmarking generalizable bimanual tool-action-object understanding. CVPR 2024.

Experiments: Results

北京大学 PEKING UNIVERSITY

- **BiDexHD-IPPO** achieves near-complete stage-one success and high tracking quality for seen objects
- **BiDexHD-IPPO+DAgger** significantly outperforms both PPO variant and BC
- Explanation for poor performance of **BC**:

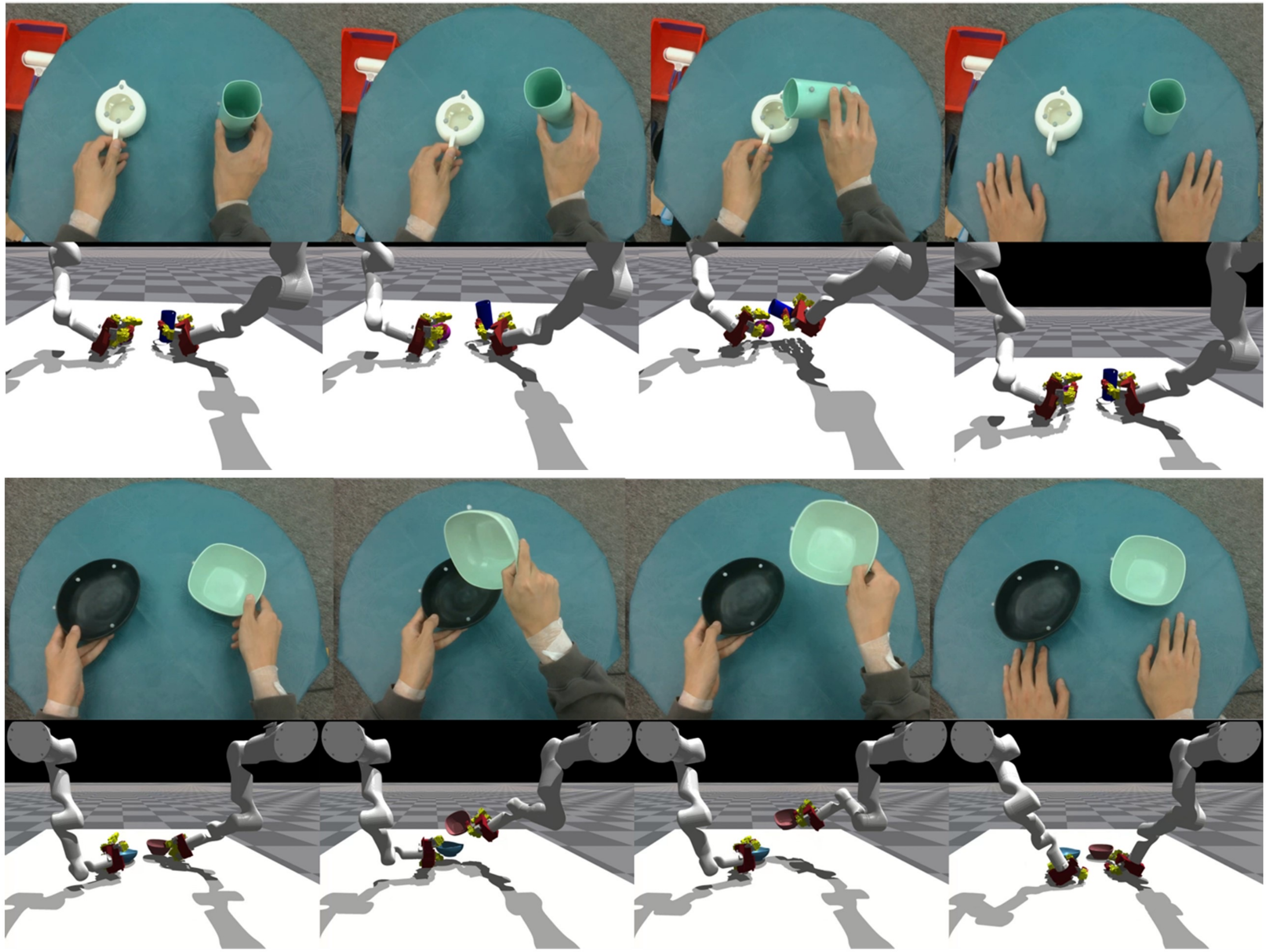
  (1) Only one available demonstration for each task (2) Lack of kinematics & dynamics

Method	Train $m_1(\%)$	Train $m_2(\%)$	Test Comb $m_1(\%)$	Test Comb $m_2(\%)$	Test New $m_1(\%)$	Test New $m_2(\%)$
BiDexHD-PPO	90.55	53.88	78.74	36.99	**81.42**	**26.24**
BiDexHD-IPPO (w/o stage-1)	25.00	17.52	24.80	18.10	19.85	08.51
BiDexHD-IPPO (w/o gc)	90.53	66.39	91.47	52.11	77.03	22.63
BiDexHD-IPPO (w/o bonus)	97.67	66.65	98.01	59.76	77.96	17.52
BiDexHD-IPPO	**98.71**	**78.18**	**98.37**	**59.94**	75.48	21.34
BC	00.00	00.00	00.00	00.00	00.00	00.00
BiDexHD-PPO+DAgger	95.35	55.82	76.75	30.42	86.34	30.00
BiDexHD-IPPO+DAgger	**99.38**	**74.59**	**92.85**	**48.43**	**94.79**	**53.71**

Experiments: Visualization

State-Based Policy Training

Vision-Based Policy Inference



Experiments: Scalability

北京大學 PEKING UNIVERSITY

ARCTIC [1] focuses on bimanual cooperative tasks of a single object.

We build up **11** tasks in total, **8** for multi-task training, **3** for testing.

Metrics (%)	BiDexHD-IPPO	BiDexHD-IPPO+DAgger
Train m_1	93.67	90.98
Train m_2	86.75	80.49
Test New m_1	80.31	88.62
Test New m_2	53.47	65.99

[1] Fan, Zicong, et al. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. CVPR 2023.

Conclusion

北京大學
PEKING UNIVERSITY

Contributions

The **three-phase** framework, BiDexHD, unifies constructing and solving tasks from human bimanual demonstrations instead of existing benchmarks, providing a **scalable** solution to diverse bimanual manipulation tasks, and paving the way for deployment.

Future Extensions

- Explore adaptive strategies to achieve more precise **spatial and temporal tracking**.
- Incorporate a wider variety of real-world tasks, such as **deformable** object manipulation and bimanual **handover**.

北京大學
PEKING UNIVERSITY

Thanks

Bohan Zhou

2026.1.22